Modelling the Risk of Exceeding Air Quality Monitoring Thresholds in Scotland

Report

October 2019

Dr Alasdair McIntosh
Dr Craig Anderson
School of Mathematics and Statistics
University of Glasgow

University of Glasgow

Scottish Government
Riaghaltas na h-Alba
gov.scot

Executive Summary

- The core of this project is to combine the air quality monitoring station data with the modelled pollution estimates produced by AEA [3] to produce enhanced estimates of the level of pollutants such as nitrogen oxides and particulate matter and also to produce estimates of the probability of them exceeding a threshold or objective level. These estimates are to be provided at a variety of spatial resolutions at the areal unit level (e.g. Data Zone) This form of modelling is known as "fusion modelling".

- A statistical model has been built based on that developed by Huang et al. [7] in order to meet these objectives. It also takes temperature and rainfall levels into account. We have adapted this model to analyse results for more than two pollutants simultaneously, taking the relationships between levels of these pollutants into account.

- Four pollutants were included in the model ($PM_{2.5}$, PM_{10}, NO_2 and NO_x). For each of these, we noted that the highest levels of pollution came in built up areas and major cities. It was noted that particulate matter levels were higher on the east coast of Scotland, probably as a result of the lower rainfall in that part of the country.

- For each pollutant, a handful of areas in the centres of cities displayed a high probability of exceeding the air quality objective levels under certain circumstances.

- A number of maps are provided in this report which illustrate some of the things users of the finished model can do:
    - Produce maps of estimated levels of a selected pollutant.
    - Produce maps of estimated risk of a pollutant exceeding a user defined level.
    - Produce maps based on user-specified weather conditions.
    - Produce maps at different geographical levels of spatial resolution such as Data Zones, Intermediate Zones, AQMAs, Council Areas or a 1km by 1km grid.

1. Introduction and Background

1.1. There has been a growth in interest in recent years in monitoring air quality, not just because of its direct effects on the environment but also because of the secondary effects that it has on human health. Indeed, a World Health Organisation report in 2016 [1] estimated that in 2012 as many as three million premature deaths were caused by outdoor air pollution. In the same year the Royal College of Physicians [2] reported that an estimated 40,000 deaths per year were due to air pollution in the UK alone.

1.2. As the policy profile of air quality has increased in recent years, a number of pieces of legislation have been aimed at regulating air quality and setting objectives for the levels of pollutants, such as the Air Quality (Scotland) Regulations 2000, the Air Quality (Scotland) Amendment Regulations 2002, Air Quality Standards (Scotland) Regulations 2010 and the Air Quality (Scotland) Amendment Regulations 2016.

1.3. Air quality monitoring stations have been set up in Scotland to monitor such pollutants as Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), Ozone (O_3) and Nitrogen oxides more generally (NO_X). Particulate matter, both solid and liquid droplets, of less that 10 microns in size (PM_{10}) and smaller ones of less than 2.5 microns in size ($PM_{2.5}$) are also measured. The air quality objective in Scotland for the annual mean level of Nitrogen Dioxide (NO_2), is 40 $\mu g/m^3$. For PM_{10}, it is 18 $\mu g/m^3$ and for $PM_{2.5}$ it is 10 $\mu g/m^3$. Objectives for SO_2 and O_3 are set at hourly levels rather than annual means; these are 350 $\mu g/m^3$ and 100 $\mu g/m^3$ respectively.

1.4. It is impractical to position a monitoring station in every location in Scotland. Therefore, modelled data are available for air quality for each square kilometre of land. These data are calculated using an atmospheric dispersion model produced by AEA [3] and available from DEFRA. While invaluable, these estimates are not as accurate as the real measurements of monitoring stations. Our goal is therefore to combine the point level measurements from monitoring stations with the grid level estimates from the AEA model, thus producing improved estimates that take all the

available information into account.

1.5. Over the past decade researchers have attempted to find ways to combine the modelled estimates with real monitoring data to produce improved estimates of air quality for areas without monitoring stations. This approach, known as “fusion modelling” is described in Fuentes et al. (2005) [4], and further extended by McMillan et al. (2010) [5] and Berrocal et al. (2010) [6]. This study will be taking such a fusion modelling approach, based on the work of Huang et al. (2016) [7].

1.6. It is therefore of interest to estimate the risk of a particular area exceeding these targets. In order to be able to make a statement about the risk of an area exceeding a monitoring threshold for a given pollutant in a given area, we must consider the uncertainty associated with our estimate. We are likely to be less certain about our estimates for regions with fewer monitoring stations, and this should be reflected when reporting our results. We present this uncertainty by producing estimates of the probability of the pollutant level exceeding the threshold.

1.7. It is also important to take the relationships between pollutants into account. Not all monitoring stations measure all the pollutants of interest. However, some pollutants are related to each other. Where, for example, NO_2 is found to be high, it might not be surprising to find NO_X is also high, thus one may not need to monitor both at each site. Huang et al. (2016) [7] also note a relationship between NO_2 and PM_{10}.

1.8. These relationships between pollutants can be incorporated into our modelling procedure, thus allowing us to improve estimates of unmeasured pollutants at and near locations where other pollutants have been measured. In other words, if a given monitoring station only measures NO_2, we can use our knowledge of the existing relationships to also estimate the levels of other pollutants (eg NO_x, PM_{10}) at this location.

1.9. The model that is developed for this report is an adaptation of the Bayesian hierarchical model described by Huang et al. (2016) [7]. Although that model was primarily motivated by modelling the health impacts of air quality, it can be usefully adapted to meet the needs of

providing predictions of the risk of exceeding air quality threshold levels. In particular, it:

- combines modelled estimates of air quality for various pollutants with actual measurements from monitoring stations;
- takes the different types of monitoring station into account;
- uses measurements for measured pollutants to make inferences about the likely levels of unmeasured pollutants;
- produces distributions of likely levels of pollutants at each location from which the risks of exceeding a particular threshold level can be derived;
- takes a series of recent historic measurements of air quality into account, increasing the amount of useful data available;
- takes mean annual temperature in each area into account.

1.10 The model required further development to deal with more than two pollutants. The Huang model only covered NO_2 and PM_{10} and only estimated rural and urban background levels of air quality. With some adjustment, the model was adapted to also estimate the risk of exceeding levels of pollutants in other environments such as at the roadside or kerbside. The model was also extended to take the effect of annual rainfall into account.

1.11 Software has been developed to allow users to easily fit this model using the statistical programming software R. This software uses a package known as RShiny to allow users to carrying out modelling and mapping of pollution data via a simple user interface. The user can thus easily map the levels of four common pollutants at multiple spatial resolutions based under a variety of modelling scenarios.

1.12 The model does not take the different heights above ground level of the monitoring station into account. According to a 2015 Scottish Government report [8], average concentrations of PM_{10} are up to 12.6% higher at 0.8m than they are at 1.68m, whereas for NO_2, where there are high ambient concentrations, higher concentrations were observed at 1.68m than at 0.8m.

1.13 The aim of this report is to provide improved estimates of the concentrations of multiple pollutants using both the monitored and modelled data, together with relevant meteorological data. The model will directly account for the correlations between different pollutants when constructing these improved estimates. Furthermore, we will improve on the AEA modelled data by also allowing for uncertainty quantification within our model. This allows risks of exceedance to be estimated in a straightforward manner using the results from our model.

2. Data and Methodology

2.1 Table 1 shows a summary of the number of monitoring stations in Scotland in each year since 2010, split by type of station and type of pollutant. Figure 1 displays a map of the locations of each of the monitoring stations, with different colours corresponding to different station types.

2.2 We can see from Table 1 that monitoring stations are more likely to be stationed at roadside locations; for example, there were 59 such stations modelling NO_2 in 2018. A key role of these stations is to identify exceedances and therefore it is unsurprising that they are located in places where the pollution levels are likely to be highest.

2.3 As a consequence, we note that there is comparatively little monitoring station data available for background urban and rural NO_2, NO_X and PM_{10} air quality. For example, there were only four rural stations monitoring NO_2 in 2018.

2.4 We also note that prior to 2015, there were only a handful of monitoring stations (a maximum of five in any year) collecting data for $PM_{2.5}$. However, such data has been more frequently collected from 2015 onwards, and there were a total of 53 stations collecting this data in 2018.

2.5 There are far fewer monitoring stations collecting data on SO_2 or O_3. In 2018 there were only nine stations gathering SO_2 levels and only 11 collecting O_3 data.

2.6 From Figure 1, we can see that the majority of the monitoring stations are located in the Central Belt of Scotland, with some others in the North-East around Dundee and Aberdeen. There are far fewer stations located in the less populated and more rural parts of the country.

2.7 In order to estimate air quality across the entire country, our approach uses the AEA [3] modelled data, which is obtained from DEFRA. For the modelled pollution data, Scotland is divided into almost 85,000 1km x 1km

grid squares, and estimation is therefore carried out on these squares.

2.8 The annual average level of rainfall and temperature for each of these grid squares is obtained from the Met Office and the classification into urban or rural is obtained from the Scottish Government's 8-fold urban rural classification. The sources for the data used can be found in the appendix.

2.9 The idea of a fusion model is to carry out a statistical regression on the monitor data using the modelled data, meteorological data and site type as potential explanatory variables. The model is structured in such a way as to allow for correlation between pollutants, thus allowing several pollutants to be modelled simultaneously. The appendix of this report contains a mathematical description of our model. Further details of the model are described in Huang et al (2016).

2.10 This model is used to predict the concentration of each pollutant in each of the 1km x 1km grid squares in Scotland. This is essentially a prediction of what the monitored concentration would be if we situated a monitoring station in that square, and is based on our knowledge of the modelled concentration, meteorological data and site type for that square.

2.11 The predicted concentrations at each of these squares are then aggregated to the desired spatial resolution (e.g. local authority, Data Zone).

2.12 This aggregation process is based on a simple weighted average based on the area of intersection between each grid square and the relevant region.

2.13 It is also possible to carry out a population-weighted aggregation which gives more weight to areas which have a higher population density. This might be appropriate if the goal if the study is to identify the extent to which the people living in a particular region are exposed to pollution.

2.14 More details on the aggregation approaches described in 2.12 and 2.13 can be found in Appendix 6.2.

2.15 The uncertainty associated with these estimates can be accounted for by the construction of an interval estimate which gives a plausible range for the pollutant concentration in each grid square or region. This uncertainty can also be used to produce an estimated probability of the pollutant exceeding a particular level, which can be particularly useful for monitoring compliance with air quality targets.

Table 1 - Air quality monitoring stations by pollutant and environment

NO_2	2010	2011	2012	2013	2014	2015	2016	2017	2018
Airport	1	1	1	1	1	0	0	0	0
Kerbside	5	4	6	6	6	6	7	7	7
Roadside	39	43	48	53	49	52	60	58	59
Rural	3	3	2	3	3	3	4	4	4
Suburban	2	1	2	2	3	3	3	3	3
Urban Background	6	6	6	4	5	5	7	7	7
Urban Centre	1	1	0	0	0	0	0	0	0
Urban Industrial	1	1	1	1	1	1	2	2	2
PM10	2010	2011	2012	2013	2014	2015	2016	2017	2018
Kerbside	3	2	4	3	3	4	6	6	6
Roadside	36	42	45	44	43	48	60	58	58
Rural	1	2	2	1	1	1	2	2	2
Suburban	0	0	0	0	1	1	1	1	1
Urban Background	9	8	4	7	6	7	8	8	8
Urban Centre	0	1	0	0	0	0	0	0	0
Urban Industrial	2	2	2	2	2	2	2	2	2
NO_X	2010	2011	2012	2013	2014	2015	2016	2017	2018
Airport	1	1	1	1	1	0	0	0	0
Kerbside	5	4	6	6	6	6	7	7	7
Roadside	39	43	48	53	49	52	60	58	59
Rural	3	3	2	3	3	3	4	4	4
Suburban	2	1	2	2	3	3	3	3	3
Urban Background	7	8	6	4	5	5	7	7	7
Urban Centre	1	1	0	0	0	0	0	0	0
Urban Industrial	1	1	1	1	1	1	2	2	2
PM2.5	2010	2011	2012	2013	2014	2015	2016	2017	2018
Kerbside	1	1	1	1	0	2	3	4	5
Roadside	0	0	0	0	0	7	17	31	40
Rural	0	1	1	0	1	1	1	1	2
Urban Background	2	2	1	2	2	3	3	4	5
Urban Centre	1	1	0	0	0	0	0	0	0
Urban Industrial	1	1	1	0	1	1	1	1	1
SO_2	2010	2011	2012	2013	2014	2015	2016	2017	2018
Kerbside	1	1	1	0	0	0	0	0	0
Roadside	3	4	4	4	2	2	2	2	2
Rural	1	0	0	0	0	0	0	1	1
Urban Background	4	4	4	4	2	3	4	3	3
Urban Centre	1	1	0	0	0	0	0	0	0
Urban Industrial	2	2	2	2	2	1	3	3	3
O_3	2010	2011	2012	2013	2014	2015	2016	2017	2018
Rural	6	6	5	5	5	6	6	6	6
Suburban	2	2	2	1	2	2	2	2	2
Urban Background	2	2	2	2	2	3	3	3	3
Urban Centre	1	1	0	0	0	0	0	0	0

Figure 1: Map of air quality monitors by type.

3. Results and Discussion

3.1. This section provides example results from the model. However, the attraction of developing a software application is that it allows the user to produce their own maps to explore different aspects of air quality that are of interest and to produce their own reports for their own purposes.

3.2. The results are most easily summarised in map form, with a colour scale to indicate pollution levels. A darker blue indicates a higher level of pollutant or a higher risk of exceeding a specified pollution level. Maps are provided in this report, but using the software the user is able to interact directly with these maps by moving the map on the screen or zooming into particular locations.

3.3. For modelling purposes, the airport, urban centre, urban industrial and suburban environment monitoring stations are grouped along with urban background ones. In the remainder of this section references to urban background monitoring stations refers to this whole group.

3.4. Figure 2 shows the NO_2 levels in the AEA modelled data as a comparison for our modelled data. Figure 3 shows the map of estimates of background NO_2 from the model. The weather conditions in each 1km grid square are assumed to be the same as the mean levels experienced over the years since 2015. These estimates are an average over for the grid square over the whole year.

3.5. We can see from Figure 3 that the highest levels of NO_2 are to be found in the cities, principally, Glasgow, Edinburgh and Aberdeen. The surrounding populated areas also experience heightened levels. The results from our model have very similar features to the AEA estimates, which is unsurprising given that it is one of the inputs into our model.

3.6. However, the results are not identical, since our fusion model allows us to use the correlation between NO_2 and the other three pollutants to improve estimation. Our model also allows the uncertainty associated with these estimates to be quantified.

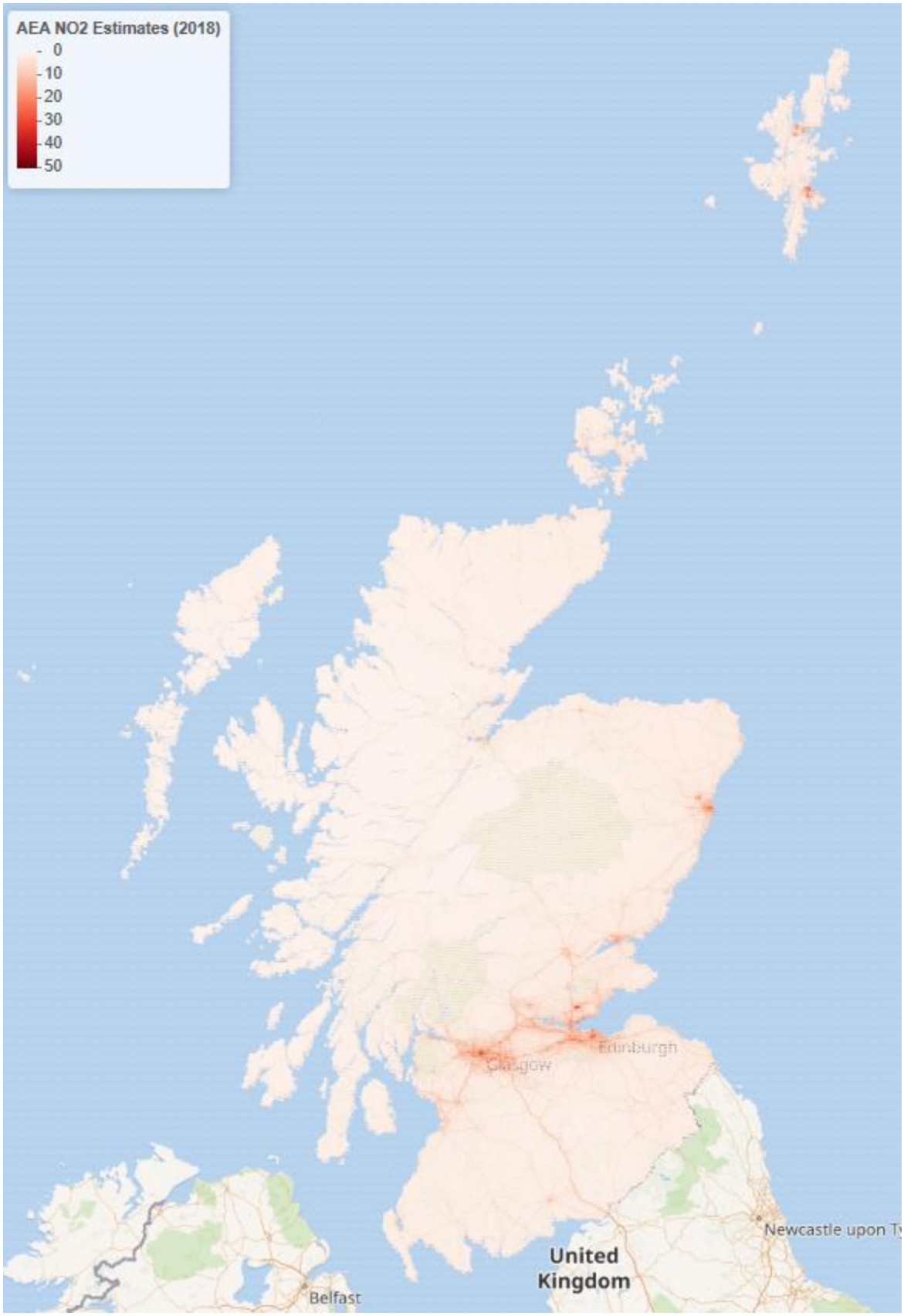

Figure 2: Levels of NO_2 based on AEA modelled data.

Figure 3: Estimated levels of background NO_2 assuming average weather conditions.

Figure 4: Estimated levels of NO_2 assuming average weather conditions. Urban locations estimated as roadside.

3.7. Roadside locations make up the lion's share of monitoring station environments. This reflects the interest and concern about air quality levels experienced by pedestrians in urban street locations. We can therefore consider modelling urban areas as though they were measured at a roadside location; thus showing the predicted pollution levels at roadside sites within each of these grid squares.

3.8. Figure 4 displays the estimated NO_2 levels from our model if we consider the urban areas as being roadside locations. We see that evaluating the air quality at the roadside in these urban areas leads to much higher estimates of NO_2 than were observed in Figure 3. Given that most of the actual monitors are found at the roadside, these estimates may more accurately reflect the observations which would be made by such monitors.

3.9. The model summarises the range of possible predicted values in each square, and thus can compute the probability of exceeding a particular threshold for each 1km square. Figures 5 and 6 illustrate the estimated probability of each area exceeding an average NO_2 concentration of 40 $\mu g/m^3$ based on the assumptions made in Figures 3 and 4.

3.10. The lack of any dark blue regions in Figure 5 suggests that the risk of exceeding the 40 $\mu g/m^3$ limit is low for the whole country based on modelling urban locations as background. However, Figure 6 shows that if urban locations are modelled as roadside, there is a high risk of exceedance in the city centres of Glasgow, Edinburgh and Aberdeen.

3.11. It is therefore important to give thorough consideration to the choice of modelling approach for urban locations. Selecting urban as background might reflect the average level of pollution people living in urban areas experience on a daily basis, while selecting urban as roadside could reflect the level of pollution experienced by pedestrians in these urban areas. The software allows the user to select the type of urban environment which is most appropriate for their own study.

Figure 5: Estimated probability of NO_2 levels exceeding the annual mean limit of 40μg/m^3.

Figure 6: Estimated probability of NO_2 levels exceeding the annual mean limit of 40μg/m^3. Urban locations estimated as roadside.

3.12. In addition to indicating the probability of exceeding an existing target, the uncertainty estimates from our model could also be used to investigate the effects of changing the objective levels for pollutants. For example, Figure 7 shows the probability of exceeding an average NO_2 concentration of 20 $\mu g/m^3$ based on urban locations being modelled as roadside. We see that many urban locations across the country would have a high probability of exceeding such a limit.

3.13. The model is also able to estimate the effects of different weather conditions on the estimated levels of each pollutant and the probabilities of exceeding the objective levels. Both rainfall and temperature have been included in our model, however rainfall is not found to be statistically significant in the case of any of the four pollutants and temperature only in the case of particulate matter. The current model does not account for wind speed, mainly because this would substantially increase the complexity (and running time) of the model, given that both speed and direction have to be accounted for.

3.14. In the case of $PM_{2.5}$, the estimated temperature effect is 0.135 with a 95% interval estimate of about (0.022, 0.275). This impacts on the natural logarithm of $PM_{2.5}$ so a 1^oC decrease in temperature would lead to a decrease in $PM_{2.5}$ level of somewhere between 2.2% and 31.7%, with our best estimate being 14.5%. The large uncertainty reflects the relatively fewer monitoring stations recording levels of this pollutant in recent years.

3.15. For PM_{10}, the estimated temperature effect is 0.096 with a 95% interval estimate of about (0.014, 0.180). Here, a 1^oC decrease in temperature would lead to a decrease in PM_{10} level of somewhere between 1.4% and 19.7%, with our best estimate being 10.1%.

3.16. These conclusions do, however need to be qualified. It may well be that the model is underestimating the effects of weather conditions because the AEA estimates of pollutants are correlated with both temperature and rainfall. The lack of significant effect of rainfall on particulate matter levels was particularly surprising and may well be a result of some of the rainfall effects already being accounted for through the AEA estimates.

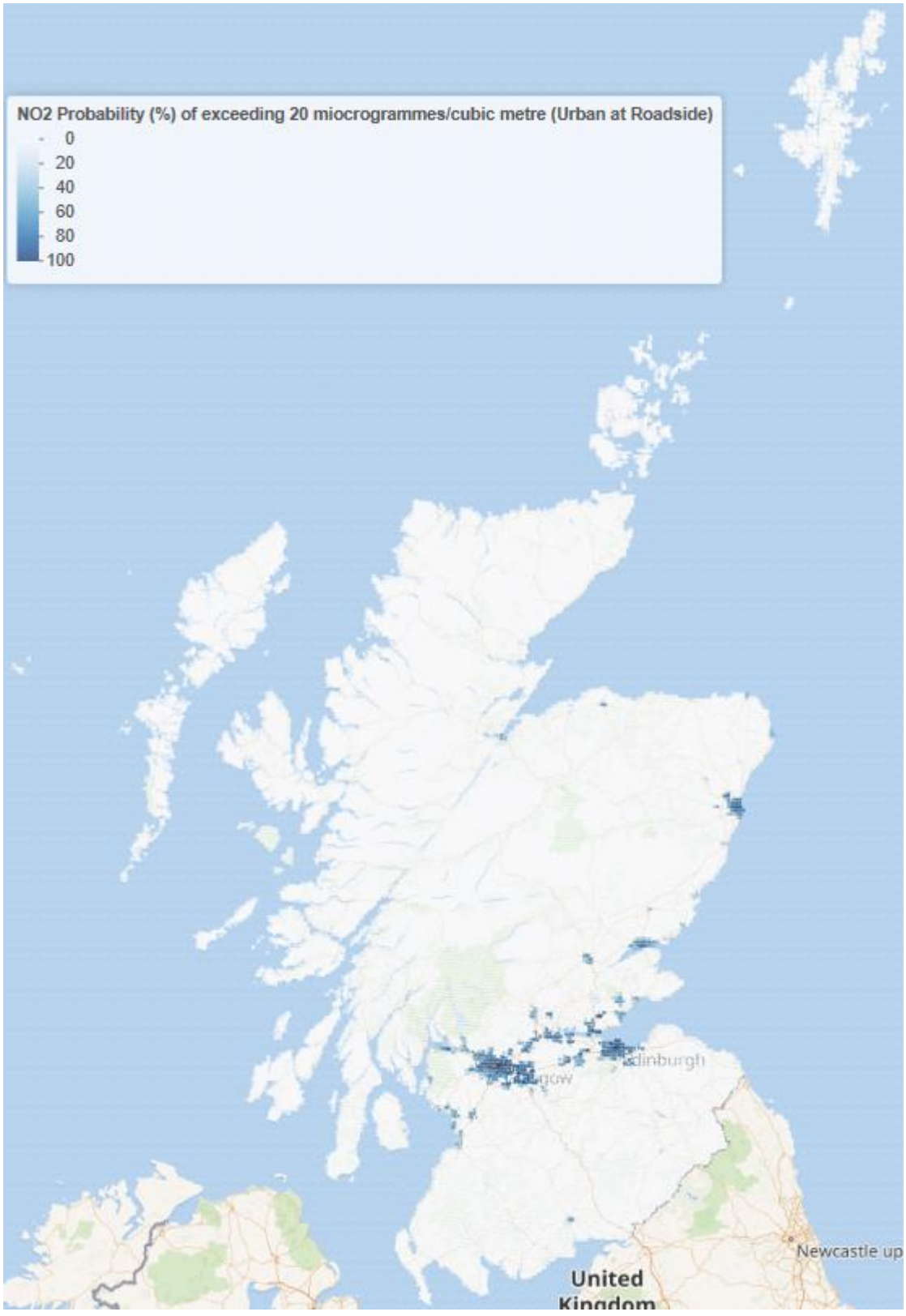

Figure 7: Estimated probability of NO_2 levels exceeding **a reduced** annual mean limit of 20μg/m^3. Urban locations estimated as roadside.

3.17. The issue can be illustrated visually by comparing a map of average annual rainfall from the Met Office (Figure 8) with a map of PM_{10} Ricardo AEA estimates (Figure 9) from the most recent year that both are available on a 1km grid scale. The driest areas in Figure 8 are very similar to the highest PM_{10} concentrations in Figure 9. The higher concentrations are in the drier areas in the east of the country, particularly, Aberdeenshire, the area around Dundee, Fife, East Lothian and a very similar feature in the eastern Borders area.

3.18. As a result, some of the variation in the data that is caused by weather conditions could be already be accounted for by the AEA estimates in our model, leading to underestimation of the effects of weather. One could draw weather effects out more effectively by building a more complex model that takes seasonal effects into account, but this would be far more computationally intensive since it would require daily/weekly weather and pollution data rather than the annual data in this model.

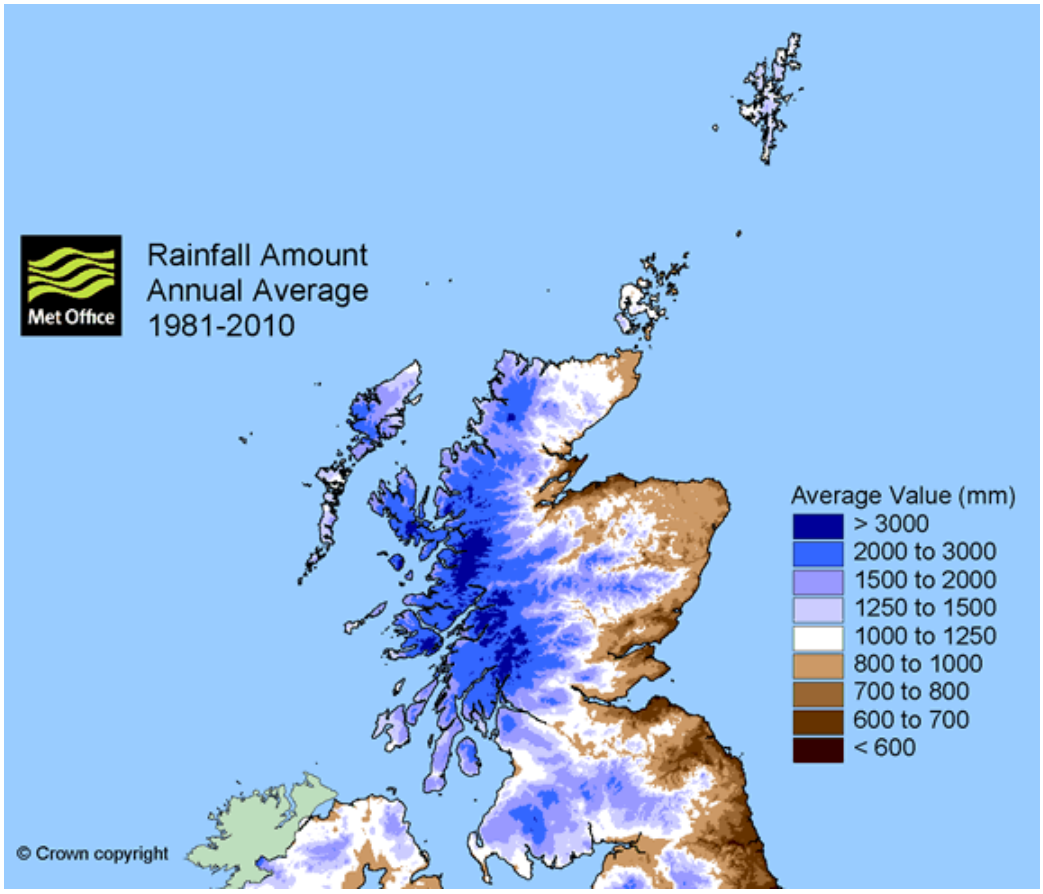

Figure 8: Map of Annual Average Rainfall 1981-2010

Note: This map has been digitally altered from the original Met Office version to remove Shetland from a box and replace it in its approximately correct geographical position.

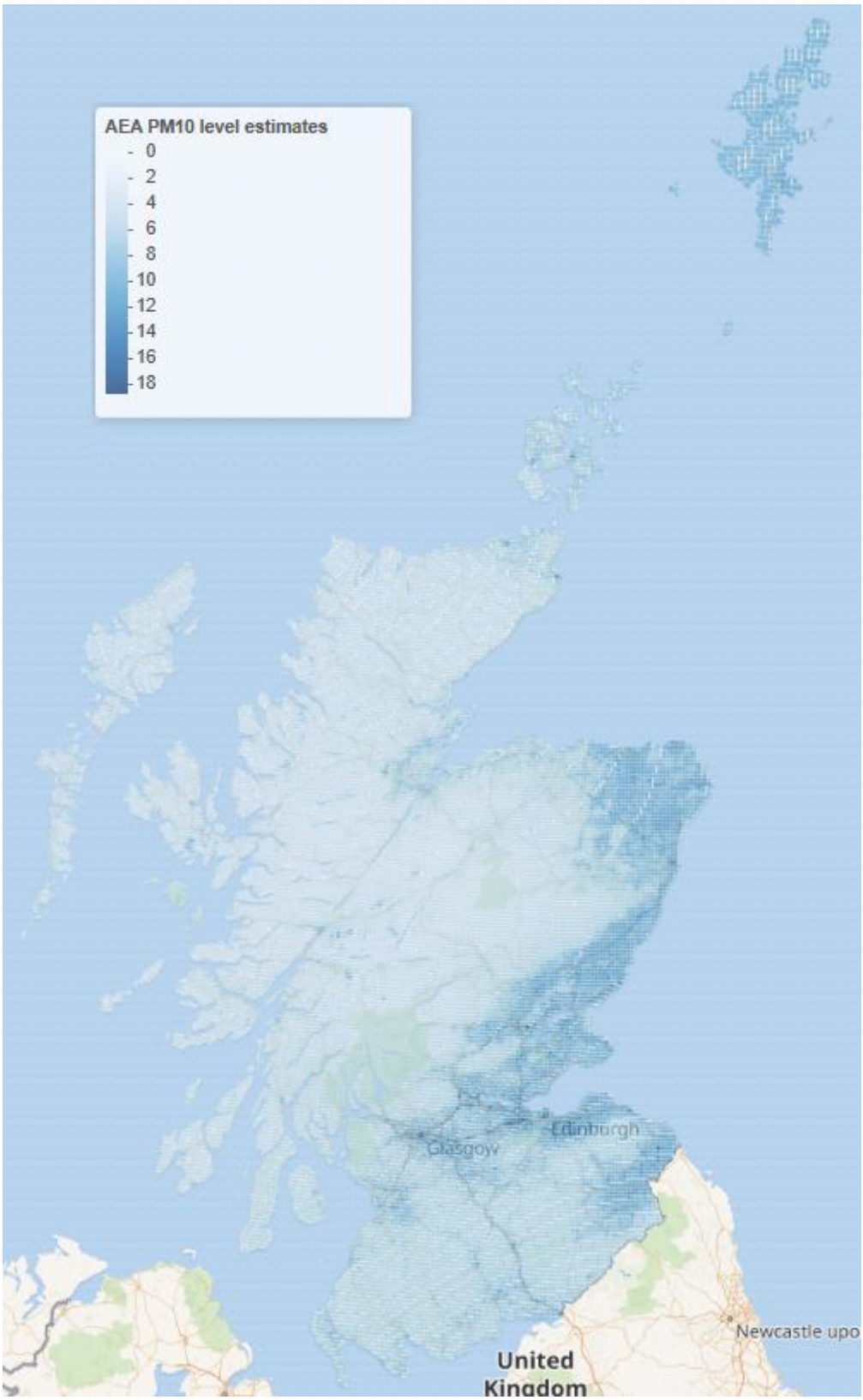

Figure 9: Levels of NO_2 based on AEA modelled data.

3.19. Any model is always only as good as the data supplied to it. The focus has been on roadside measurements in recent years with monitoring stations sited, for understandable reasons, in areas where air quality may be a particular concern. However, these may not represent air quality over the whole country or indeed even within the whole square kilometre in which they are sited.

3.20. Increasing the low sample sizes of other types of monitoring station particularly rural ones but also urban background ones in representative areas would assist with the accuracy of any model of this type. The small sample sizes could make it difficult for the model to establish an accurate baseline difference in the levels of pollutant between the two types of site.

3.21. Figure 10 shows the background distribution of PM_{10} across Scotland as estimated by the model. The map also shows an example of a map at Data Zone level. Higher levels of the pollutant are found in and around built up areas, but the east coast of the country also experiences higher levels than the west, particularly in Aberdeenshire, Fife, East Lothian and the area around Dundee.

3.22. Figure 11 shows the distribution of NO_X, also at background in a Data Zone map. The distribution is highly concentrated in major centres of population, and unlike PM_{10} there does not appear to be any evidence that levels are higher on the east coast of the country.

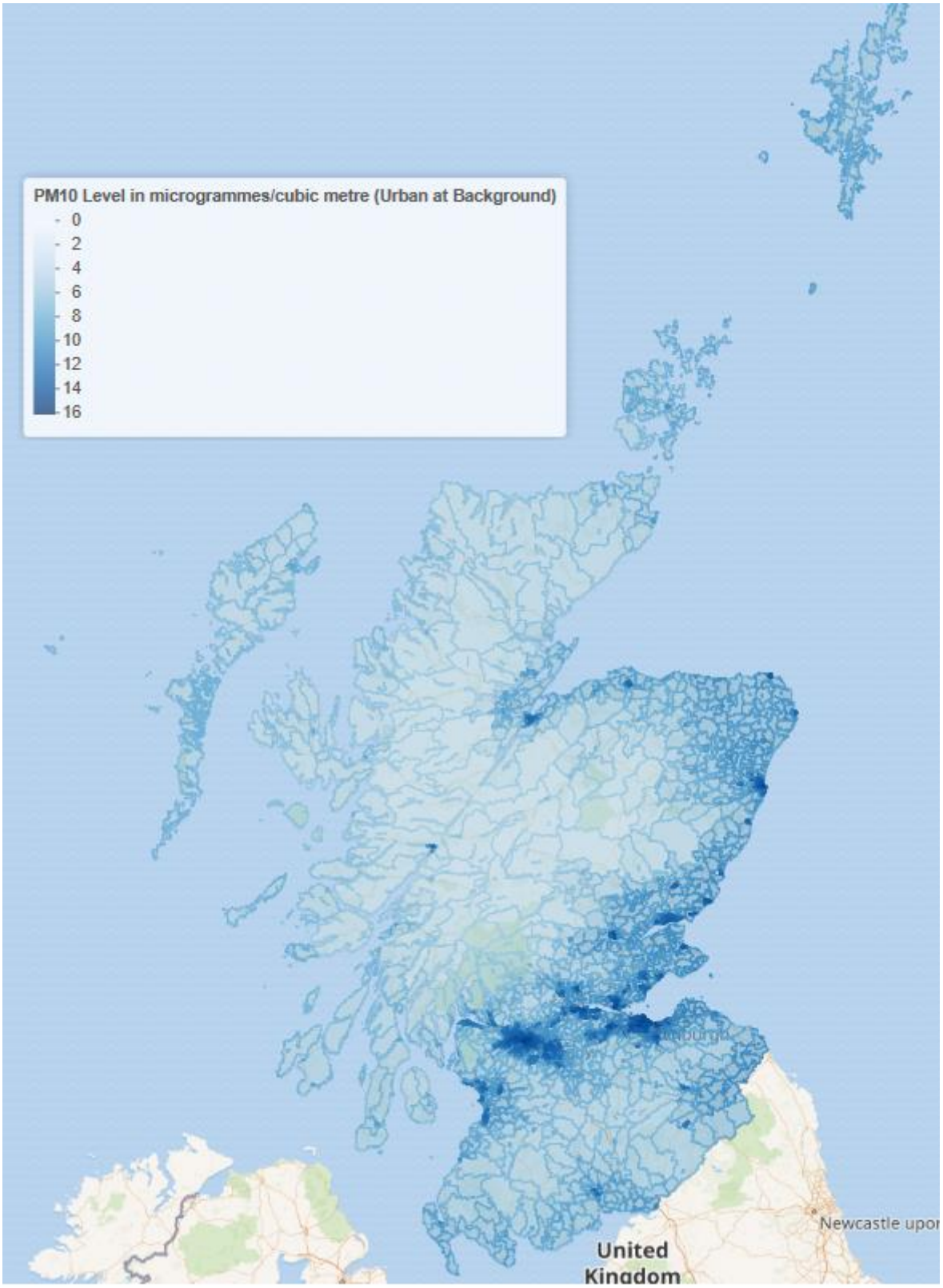

Figure 10: Estimated levels of background PM_{10} assuming average weather conditions by Data Zone.

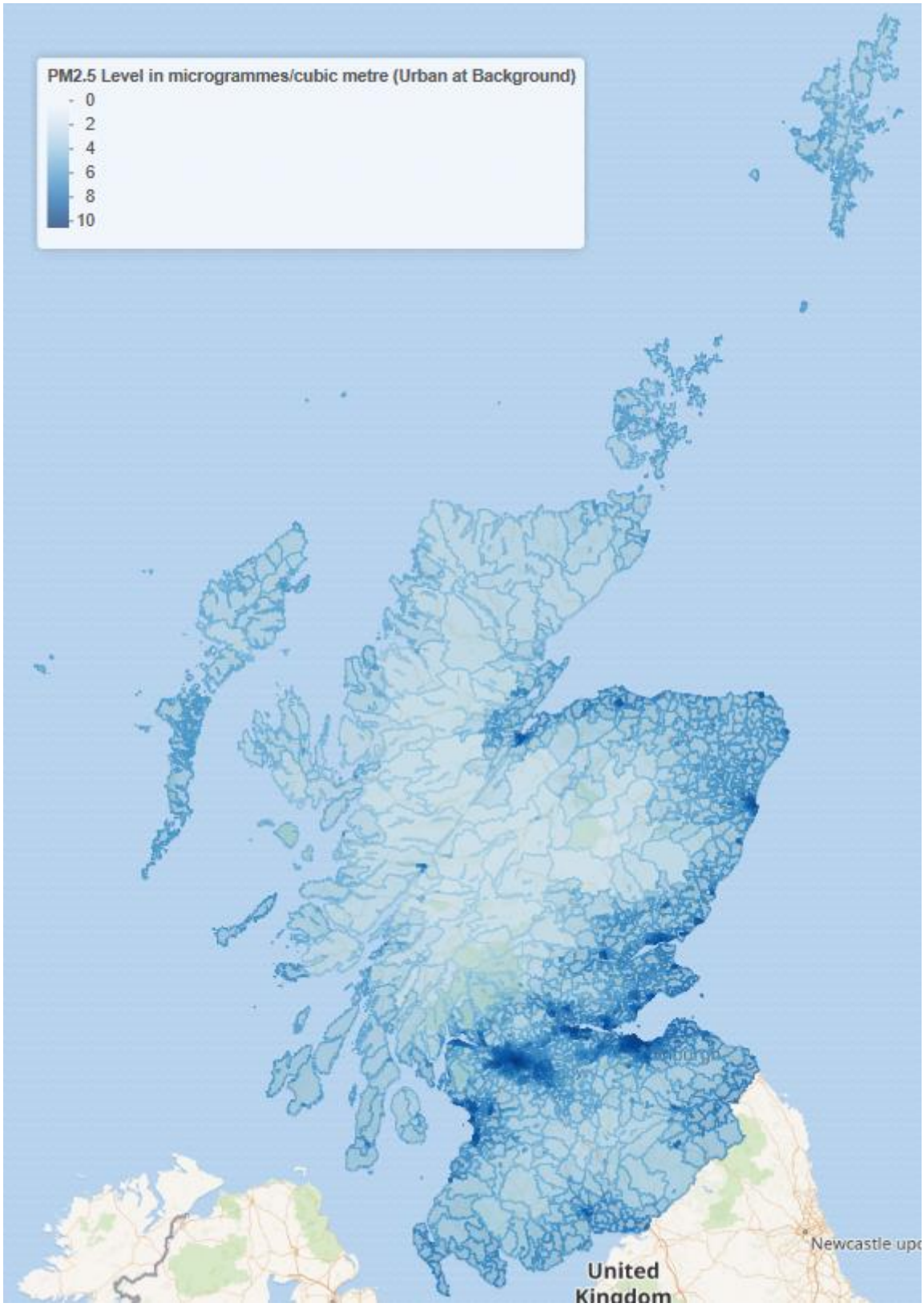

Figure 11: Estimated levels of background $PM_{2.5}$ assuming average weather conditions by Data Zone.

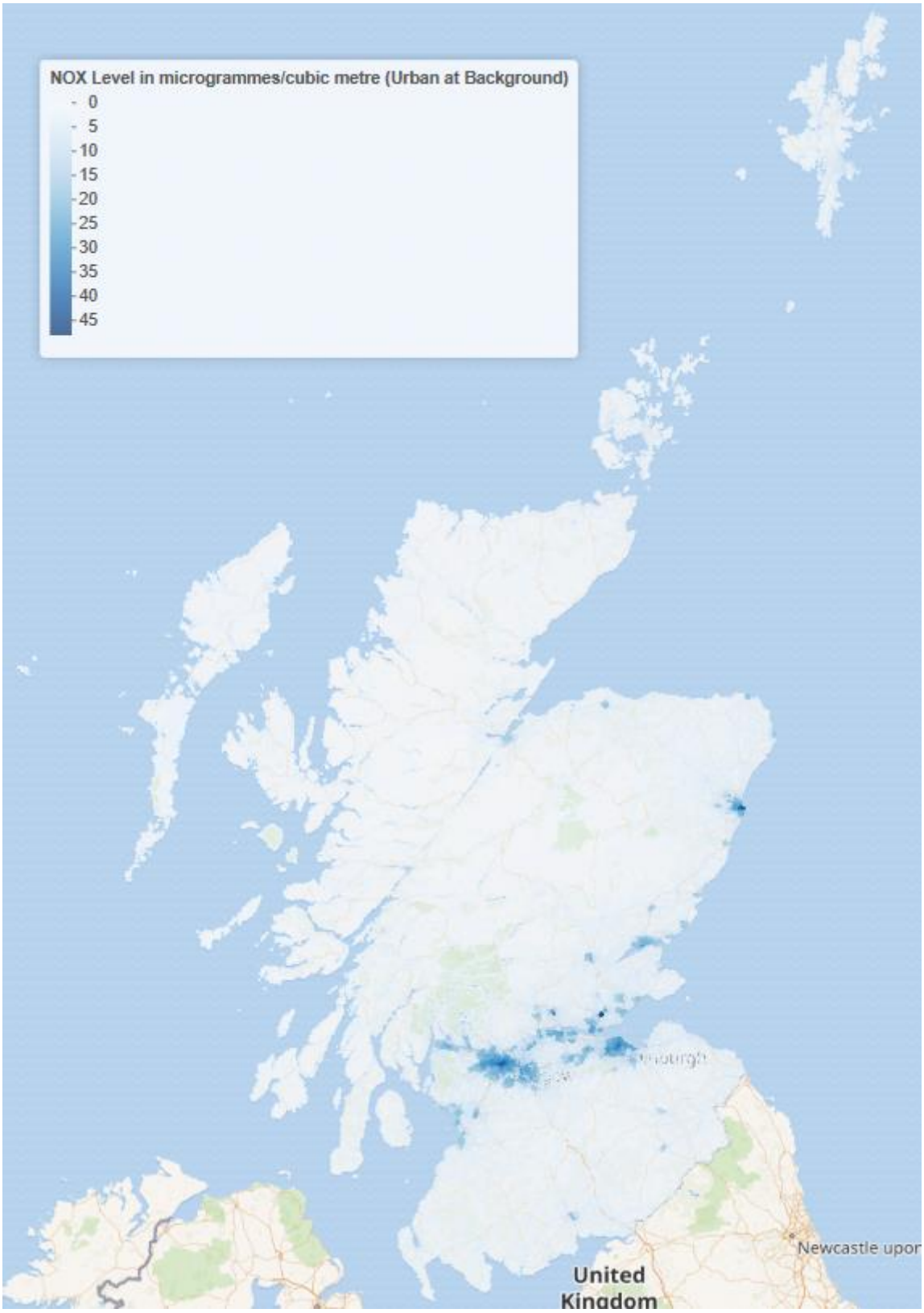

Figure 12: Estimated levels of background NO_X assuming average weather conditions by Data Zone.

4. Possible Future Work

4.1. There are a number of possible enhancements that could be made to the model in future.

4.2. A model could be developed that takes seasonal variation into account. At present the model is based on annual averages and totals. Variation can also happen within years. It has been noted that the present model may be underestimating the effects of weather on pollutant levels. This may be because of the similarity between patterns of weather and the AEA estimates of pollutant levels. Having a model where weather changes within a year could help to draw out the effects of weather and result in a model that is valid for a wider range of weather variation than is possible with the present one. However, it is likely that such a model would be much more computationally intensive.

4.3. Monitoring data for O_3 and SO_2 levels is starting to be gathered. At present, there is not enough data to include these pollutants in the model in an effective way. In the next few years, however, this is likely to change. In principle, these pollutants could be incorporated into the model. This would allow us to estimate the levels of these pollutants in the same way, and could also potentially improve estimation of the existing pollutants since additional between-pollutant correlations could be incorporated into the model.

4.4. Other weather and environmental conditions such as wind speed, wind direction and sunlight could be considered within the model. The latter would be particularly relevant if O_3 is monitored.

4.5. The model could be enhanced to take the height of the monitoring station from the ground into account. A Scottish Government report [8] showed that distance from the ground could affect pollution levels. Including the height of the monitoring stations would lead to more precise estimation of pollution levels. Additionally, this would allow pollution maps to be produced for different heights, which could, for example, explore the differences in pollution levels experienced by adults and children.

Similarly, the street canyon affect could be taken into consideration here.

4.6. At present, the navigation of the pollution maps in the software (eg moving around, zooming) can be slightly sluggish on some computers, particularly when there are a lot of shapes drawn on the map. This is a particular problem in the 1km square grid maps. The maps are drawn using a piece of software called Leaflet, and an updated version of this software called LeafGL is in development. This updated version will draw the map polygons much faster in real time. When it becomes ready, it may be possible to update our application to use it.

4.7. A fusion model of the type outlined here incorporates pollution data from multiple sources; in our case, monitoring data and modelled AEA estimates. The model also allows different levels of uncertainty to be associated with different data sources. This approach could be extended to include further pollution data from additional sources; for example hyperlocal data collected by vehicles, diffusion tubes or smaller, cheaper household monitors. Hyperlocal monitoring is being rolled out in London [9], and a model such as ours would allow data collected from such sources to be incorporated into the estimates, whilst taking into account the increased uncertainty associated with these smaller monitors compared to the existing, more reliable fixed monitors.

4.8. Fusion models have also been used to estimate the effects of air pollution on health [7]. Our model could be extended to include health data (eg respiratory hospital admissions) in order to quantify the effects of air pollution on health in Scotland, and to identify the potential health impacts of changes to air quality policy in Scotland.

5. Additional Material

5.1. This report is accompanied by software which allows users to easily fit this model based on a variety of specifications, and produce maps for all four pollutants at multiple spatial resolutions.

5.2. This software is based on the using the statistical programming software R and uses a package known as RShiny to produce a simple user interface.

5.3. Detailed instructions on how to install and operate the software are provided in the accompanying manual.

5.4. The report is also accompanied by several maps displaying the estimated pollution levels and probabilities of exceedance for a selection of pollutants at a selection of spatial resolutions. Further maps can be generated using the software.

5.5. These maps are accompanied by a map index which provides a detailed outline of the conditions under which each of the maps were generated.

5.6. The underlying raw data associated with each of these maps is also provided in a series of .csv files. These can be opened using Excel, R and many other software packages.

References

1. WHO. “Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease”. *Geneva: World Health Organisation*; 2016.

2. RCP. “Every Breath We Take: The Lifelong Impact of Air Pollution”. London: *Royal College of Physicians*; 2016.

3. AEA. “UK Modelling Under the Air Quality Directive (2008/50/EC) for 2010 Covering the Following Air Quality Pollutants SO2, NOx, NO2, PM10, PM2.5, Lead, Benzene, CO and Ozone”; 2011.

4. Fuentes M, Raftery AE. “Model Evaluation and Spatial Interpolation by Bayesian Combination of Observations with Outputs from Numerical Models”. *Biometrics*. 2005;61(1): 36-45.

5. McMillan NJ, Holland DM, Morara M, Feng J. “Combining Numerical Model Output and Particulate Data using Bayesian Space-Time Modeling”. *Environmetrics*. 2010;21(1): 48-65.

6. Berrocal VJ, Gelfand AE, Holland DM. “Spatio-Temporal Downscaler for Output from Numerical Models”. *Journal of Agricultural Biological and Environmental Statistics* 2010;15(2): 176-197.

7. Huang G, Lee D, Scott EM. “Multivariate Space-Time Modelling of Multiple Air Pollutants and their Health Effects Accounting for Exposure Uncertainty”. *Statistics in Medicine* 2018;37(7): 1134-1148.

8. The Scottish Government. “Air quality study: assessing variations in roadside air quality with sampling height”. 2015.

9. The Mayor of London. “Mayor launches world’s largest air quality monitoring network”. 2019.

6. Appendices

6.1 Mathematical Description of the Model

6.1.1 The n measured pollution levels at monitoring sites are at locations $(s_1,\ldots,s_n)$, for q pollutants for year t are denoted as $(\boldsymbol{X}_{1t},\ldots,\boldsymbol{X}_{qt})$, where $\boldsymbol{X}_{jt}=\left(X_{jt}(s_1),\ldots,X_{jt}(s_n)\right)$, the set of observations over all monitoring stations for pollutant j in year t. These are modelled as,

$$\begin{bmatrix} \boldsymbol{X}_{1t} \\ \boldsymbol{X}_{2t} \\ \ldots \\ \boldsymbol{X}_{qt} \end{bmatrix} \sim N\left[\begin{bmatrix} \boldsymbol{Z}_{1t} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Z}_{qt} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{1t} \\ \ldots \\ \boldsymbol{\beta}_{qt} \end{bmatrix}, \sigma^2_t \boldsymbol{C}_{qxq} \otimes \boldsymbol{I}_n \right], t=1,\ldots,T,$$

which is a linear regression for with means $(\boldsymbol{Z}_{1t}\,\boldsymbol{\beta}_{1t,\ldots,}\,\boldsymbol{Z}_{qt}\,\boldsymbol{\beta}_{qt})$ for the q pollutants in year t. $(\boldsymbol{Z}_{1t,\ldots,}\,\boldsymbol{Z}_{qt})$ represents the n x q design matrices. These include an intercept term, a factor for the type of monitoring station, the modelled air quality data for the pollutants, annual average temperature and annual average rainfall. The corresponding regression parameters for year t, $(\boldsymbol{\beta}_{1t,\ldots,}\,\boldsymbol{\beta}_{qt})$ are vectors of length p where p is the number of parameters in the model.

6.1.2 The $\boldsymbol{\beta}_t$ are assumed to be autocorrelated in time and to follow a centred first order autoregressive process. In the case of t=1,

$$\begin{bmatrix} \boldsymbol{\beta}_{11} \\ \ldots \\ \boldsymbol{\beta}_{q1} \end{bmatrix} \sim N\left[\begin{bmatrix} \boldsymbol{\beta}_1 \\ \ldots \\ \boldsymbol{\beta}_q \end{bmatrix}, \tau^2 \boldsymbol{I}_{pqxpq} \right],$$

and in the other cases,

$$\begin{bmatrix} \boldsymbol{\beta}_{11} \\ \ldots \\ \boldsymbol{\beta}_{q1} \end{bmatrix} \sim N\left[\begin{bmatrix} \boldsymbol{\beta}_1 + \kappa(\boldsymbol{\beta}_{1(t-1)} - \boldsymbol{\beta}_1) \\ \ldots \\ \boldsymbol{\beta}_q + \kappa(\boldsymbol{\beta}_{q(t-1)} - \boldsymbol{\beta}_q) \end{bmatrix}, \tau^2 \boldsymbol{I}_{pqxpq} \right], t=2,\ldots,T,$$

where,

$$\begin{bmatrix} \boldsymbol{\beta}_1 \\ \dots \\ \boldsymbol{\beta}_q \end{bmatrix} \sim N \left[\begin{bmatrix} \mathbf{0} \\ \dots \\ \mathbf{0} \end{bmatrix} , 1000\boldsymbol{I}_{pqxpq} \right] .$$

6.1.3 The extent of the temporal autocorrelation is governed by κ which is given a Uniform[0,1] prior. When κ=0, the $\boldsymbol{\beta}_{jt}$ are estimated independently for each year and so the ($\boldsymbol{\beta}_{1t,\dots,}\ \boldsymbol{\beta}_{qt}$) are smoothed towards overall means ($\boldsymbol{\beta}_{1,\dots,}\ \boldsymbol{\beta}_{q}$) respectively. In the other extreme case of κ=1, the ($\boldsymbol{\beta}_{1t,\dots,}\ \boldsymbol{\beta}_{qt}$) are maximally correlated with the parameters for the previous year, ($\boldsymbol{\beta}_{1(t\text{-}1),\dots,}\ \boldsymbol{\beta}_{q(t\text{-}1)}$).

6.1.4 The covariance matrix, $\sigma^2{}_t\ \boldsymbol{C}_{qxq} \otimes \boldsymbol{I}_n$ assumes correlations between pollutants at each monitoring site but not across sites. $\boldsymbol{I}_n$ is an nxn identity matrix where n is the number of monitoring stations. $\boldsymbol{C}_{qxq}$ is the matrix for the covariance between pollutants at the same site so that C_{ij} represents the covariance between pollutants i and j at the same site. This covariance matrix is assumed to follow an inverse Wishart distribution, $\boldsymbol{C}_{qxq} \sim \text{Inverse-Wishart}(q, 100\ \boldsymbol{I}_q)$. The scaling parameter, $\sigma^2{}_t$ allows for different levels of residual variation over time. It is assumed to be temporally autocorrelated using a first order random walk prior. It must be non-negative. To prevent it becoming negative, the log scale is used.

$$\ln(\sigma^2{}_t) \sim N\left(\ln(\sigma^2{}_{t\text{-}1}), \sigma^2\right), \ t=2,\dots,T,$$

$$f\left(\ln(\sigma^2{}_t)\right) \propto 1.$$

6.1.5 The parameters σ^2 and τ^2 have weakly informative prior distributions and are assumed to be Inverse-Gamma distributed,

$$\sigma^2 \sim \text{Inverse-Gamma}(0.001, 0.001),$$
$$\tau^2 \sim \text{Inverse-Gamma}(0.001, 0.001).$$

6.2 Aggregation of grid data

6.2.1 From our model, we obtain a set of pollution estimates at the grid level (G) and we wish to aggregate these to the region level (R) to reflect the average pollution level across the region.

6.2.2 Let $Y(G_i)$ be the estimated level of a pollutant in grid square G_i, and let $A(R_j \cap G_i)$ be the area of intersection (in km2) between grid square G_i and region R_j.

6.2.3 Then we can estimate the pollutant level in region $R_j$using the formula:

$$Y(R_j) = \sum_{i=1}^{m} \frac{A(R_j \cap G_i)}{\sum_{k=1}^{n} A(R_k \cap G_i)} Y(G_i)$$

so that the pollutant rates are averaged based on the proportion of the grid square which lies within the region. This is a simple weighted average based on the area of intersection between each grid square G_i and region R_j.

6.2.4 In some cases we may instead prefer an aggregation which is weighted based on the populations of the grid squares. This would ensure that more weight is given to squares where more people live, thus giving a more appropriate estimation of the average pollutant level experienced by a person living in the region. We can extend the approach outlined in 6.2.3 above to carry out this form of weighting.

6.2.5 Let $P(R_j \cap G_i)$be the population within the area of intersection $A(R_j \cap G_i)$. Then the population-weighted estimate takes the form:

$$Y(R_j) = \sum_{i=1}^{m} \frac{P(R_j \cap G_i)}{\sum_{k=1}^{m} P(R_j \cap G_k)} Y(G_i)$$

6.2.6 However, in practice we do not know the value of $P(R_j \cap G_i)$, so we must make the assumption of an equal population density across each grid square. This gives us a final estimate of the form:

$$Y(R_j) = \sum_{i=1}^{m} \frac{P(G_i)\dfrac{A(R_j \cap G_i)}{\sum_{q=1}^{n} A(R_q \cap G_i)}}{\sum_{k=1}^{m} P(G_k)\dfrac{A(R_j \cap G_k)}{\sum_{q=1}^{n} A(R_q \cap G_k)}} Y(G_i)$$

6.2 Data Sources

6.2.1 Information on mean annual temperature and total rainfall was obtained from the Met Office's HadUK-Grid dataset. It provides interpolated information on a 1km x 1km grid. It can be found at https://www.metoffice.gov.uk/climate/uk/data/haduk-grid/datasets. 2018 data is not yet available. For now, this was estimated by calculating the annual average change in temperature and rainfall for Scotland and applying those changes to the 2017 data.

6.2.2 The measured air quality data was obtained from Air Quality in Scotland and can be found at http://www.scottishairquality.scot/data/data-selector.

6.2.3 Modelled air quality data for 2015-2018 for NO2, NOx and PM10 was found at Data for Local Authority Review and Assessment purposes in the Air Quality in Scotland website at http://www.scottishairquality.scot/data/mapping?view=data.

6.2.4 Information for these pollutants for the years 2010-2014 and for PM2.5 for 2010-2018 were downloaded from https://uk-air.defra.gov.uk/data/laqm-background-maps?year=2015

6.2.5 Gridded population data on the 1km x 1km scale were obtained from the Centre for Ecology & Hydrology (CEH). https://catalogue.ceh.ac.uk/documents/0995e94d-6d42-40c1-8ed4-5090d82471e1

6.2.6 Information on which areas are classed as urban and which are rural was obtained from the Scottish Government's website at https://www2.gov.scot/Publications/2018/03/6040/downloads. The information was obtained at postcode level and the postcodes converted into latitude and longitude grid references using an online batch converter found at https://gridreferencefinder.com/postcodeBatchConverter/

6.2.7 Different datasets used different 1km x 1km grids. Where there was not an exact match, the grid for NO2 modelled air quality from Air Quality in Scotland was used as a base and each observation in each other dataset matched to the nearest point on that grid.